DOCUMENT RESUME

ED 449 210 TM 032 355

AUTHOR Long, James
TITLE An Introduction to and Generalization of the "Fail-Safe N."
PUB DATE 2001-02-01
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Meta Analysis; *Statistical Significance
IDENTIFIERS *Fail Safe Strategies

ABSTRACT

The fail-safe N is typically a "what if" analysis applied to studies rather than to a single study. This statistical procedure provides information regarding the stability of a meta-analysis by demonstrating how many nil-null articles would be needed to change the statistically significant results to a statistically nonsignificant finding. The relevant issues such as statistical significance testing, the "file drawer problem," and how they relate to the fail-safe N are discussed. The paper also explains the fail-safe N procedure in some detail using concrete heuristic examples, and a generalization of the method for use in a single study. (Contains 2 tables and 17 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

RUNNING HEAD: Fail-Safe N

An Introduction to and Generalization of the "Fail-Safe N"

James Long

Texas A & M University 77843-4225

ED 449 210

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Long
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1, 2001.

TM032355

BEST COPY AVAILABLE

Abstract

The fail-safe N is typically a "what if" analysis applied to studies rather than to a single study. This statistical procedure provides information regarding the stability of a meta-analysis by demonstrating how many nil-null articles would be needed to change the statistically significant results to a statistically non-significant finding. The relevant issues such as statistical significance testing, the "file-drawer problem", and how they relate to the fail-safe N are discussed. The paper also explains the "fail-safe N" procedure in some detail using concrete heuristic examples, and a generalization of the method for use in a single study.

An Introduction to and Generalization of the "Fail-Safe N"

Research integration has become a popular method of examining hypotheses in the realm of social science and education in the recent past. Glass (1976) first coined the term "meta-analysis" to describe the statistical method of combining various results from independent studies on similar topics and integrating the findings to provide a more global description of the variable being measured. Glass defined this procedure as "the analysis of analyses" (Glass, 1976). Since that article, many statisticians and researchers have written on the pros and cons of performing meta-analyses.

Eysenck (1978) referred to the statistical procedure as "mega-silliness." He stated that the inclusion of different types of statistical methods and varying degrees of design soundness would inevitably have a negative result on the empirical integration of diverse studies. However, since Eysenck's article many statisticians including Brown (1992), Orwin (1983) and Rosenthal (1979) have described the many positive aspects of the meta-analytic method.

One of the main complications in using this particular technique is the "file-drawer problem" that has plagued experimental studies in the social sciences since the inception of research journals. Rosenthal (1979) described this phenomenon as follows:

> ... "the file drawer problem," is that the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > .05$) results. (Rosenthal, 1979 p. 638)

The problems surrounding the use of significance tests to determine the worth of research articles have been well chronicled (Cohen, 1994; Greenwald, 1975; Thompson, 1989, 1996). Variables such as sample size, the alpha level set by the researcher, and the failure to report effect sizes in research articles have all contributed to the criticism of traditional significance testing. Across both decades and diverse disciplines these criticisms have been mounted with exponentially increasing frequency (Anderson, Burnham, & Thompson, 1999).

This realization has led to an increased emphasis on result replicability as opposed to result statistical significance, given that statistical significance does not evaluate result replicability (Cohen, 1994; Thompson, 1996). The emphasis on replicability is reflected in the recent report of the APA Task Force on Statistical Inference:

> We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses... Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies (Schmidt, 1996). (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599)

Because the literature has been biased in favor of statistically significant results (Rosenthal, 1979), such Type I errors are afforded priority for publication, but the

replications with statistically non-significant results will compete at a disadvantage for journal space, and so the self-correction of science through replication is impeded.

This problem has a particularly strong effect on the statistical procedures of a meta-analysis. Because a meta-analysis, by definition, is a synthesis of the available literature on a particular topic or variable, the influence that statistical significance testing has on the availability of only articles that produced significant findings is detrimental to the very essence of this procedure.

In order to combat this problem, Rosenthal developed the fail-safe N statistic. As Brown (1992) states, "The fail-safe N statistic is a follow-up test used with meta-analysis to estimate the number of new, unpublished, or unretrieved nonsignificant (null-result) studies that would, on the average, change the significance of a meta-analysis study to nonsignificance" (p. 179). This reasoning was what led Rosenthal to develop the fail-safe N procedure to determine exactly how many studies would be needed to change meta-analytic results from significant to non-significant.

Fail-Safe N

The fail-safe N procedure can be viewed as a "what if" analysis applied to studies rather than to a single study, as is the case with most "what if" analyses (Thompson & Kieffer, 2000). The main premise of these analyses is the fact that when "nil" null hypotheses are used, the null will always be rejected at some sample size. As Hays (1981) emphasized, "virtually any study can be made to show significant results if one uses enough subjects" (p. 293).

To combat this issue, Thompson (1996) emphasized the need to use effect sizes in the reporting and interpretation of research studies. Doing this allows the reader and the

researcher to gain a better understanding of the results instead of viewing the research material through the sometime distorted lens of traditional significance testing. Through the use of the "what if" analysis, the researcher is able to provide information concerning the exact sample size that would be necessary to produce statistically significant or nonsignificant results within the study.

Thompson and Kieffer (2000) also state that this method "helps researchers interpret their results by considering the extent to which sample size (as against effect size) yielded statistical significance" (p. 6). By teasing out the methodological factors that could have a detrimental effect on a particular study, the researchers can have more faith in reporting their methods and results in studies.

When the "what if" procedure is applied to meta-analysis studies, such as in the fail safe N, the researcher is able to give a better perspective as to the stability of the results. Brown (1992) defines stability as "the degree to which significant or insignificant empirical samples used to perform meta-analysis would change the results" (p. 180). A study is interpreted as more stable as the number of studies needed to reject statistical significance increases within the fail-safe N calculations. On the opposite side, a study is considered to be less stable as the number of studies needed to reject statistical significance decreases within the calculations.

Simplified, this means that if a fail-safe N analysis finds that 2 studies with a null-effect are needed to change the results from statistically significant to nonsignificant, it is apparent that in this particular instance the results should be interpreted with caution. This reservation is due to the practice of biased reporting of statistically significance in journal articles that was presented earlier in the paper. If only two articles reporting a

null-effect are required to change statistical significance, it could conceivably be possible that these articles exist but were not published due to the fact that they did not attain statistical significance. In turn, because the studies involving nonsignificance were not publicized, they could have been easily overlooked even in a thorough search of the literature and therefore not included in the meta-analysis.

If the fail-safe N analysis reveals that 45,357 articles reporting a null-effect would be needed to reverse a finding of statistical significance, these results could be presented with some confidence. This would be due to the unlikely possibility that this many articles actually exist and were possibly overlooked in the search of the literature.

Although there are no firm guidelines concerning the number of fail-safe N studies needed to determine true stability in research findings, there have been some speculations. Rosenthal (1969) wrote "... one could regard as resistant to the file drawer problem any combined results for which the tolerance level X reaches 5k + 10" (p. 640), where k equals the number of studies in the meta-analysis. This means that if a meta-analysis contained 40 studies, if the fail-safe N value were 210 then the results could be considered stable. However, within this same example, as the number of fail-safe N studies declines the amount of confidence that can be put into this set of statistical results also decreases.

For these reasons, researchers utilizing the meta-analytic method have been urged to perform these analyses to provide more information in the interpretation of their results. This is demonstrated in the research reported by Carson, Schriesheim and Kinicki (1990). They applied the procedure in three meta-analytic applications and found that "calculation of a fail-safe N may have led to more cautious and circumspect

interpretations of previous meta-analytic results" (p. 233). This helps to solidify the necessity of the fail-safe N statistic in meta-analytic research.

Two Methodological Examples of the Fail-Safe N

Two different statistical formulas have been introduced to calculate the fail-safe N in meta-analyses. Rosenthal (1979) provided a formula that used the combined Z-scores from the articles included in the meta-analysis to determine the number of null-effect studies necessary to reverse the statistical significance of a particular meta-analytic study. This ground- breaking procedure was the first to open the door to Orwin, who later developed another method to examine the data produced by meta-analysis. Orwin (1983) found that by using effect sizes instead of Z scores, it was possible to determine the necessary number of null-effect studies necessary to change a statistically significant finding to one that was not statistically significant. Both of these methods are still used today and each offers unique benefits in its approach. These two methods are discussed in the following paragraphs with examples to help demonstrate these points.

Rosenthal's Fail-Safe N Using Combined Z Scores

Rosenthal (1979) first developed the fail-safe N procedure in 1979 using the sum of the Z scores from the meta-analysis in the equation. The formula is as follows:

$$X = [(SUM\ Z)^2 / G] - k$$

Where X = the number of studies needed to reverse the statistically significant findings
k = the number of studies combined in the meta-analysis
(SUM Z) = the sum of the Z scores for the individual studies
G = the Z-score that falls at the p-critical value being evaluated

Although this formula is still popular today, one of the drawbacks to using this method is that its applicability is limited to use with probability levels. The researcher

can substitute the p-critical value into the formula that he or she wishes to solve. For example, if the researcher wanted to set the p-critical value at .05, a Z-score of 1.645 would be found in the denominator of the formula. However, if the researcher wanted to be more conservative in the research, the p-critical value could be set at .01, which would coincide with a Z-score of 2.33. Although these two p-critical values are the most often used, the researcher can substitute any value that is desired into the formula and discover the number of nil-null studies that are needed to reverse statistically significant meta-analytic findings.

As in all studies using the p-critical value, the actual value chosen has a great effect on the perceived stability of the meta-analytic study. If the researcher chooses to use a p-critical value of .01, then a smaller number of actual studies involving nil-null results would be needed to reverse the findings of statistical significance as opposed to a p-critical value of .05 in the same meta-analysis. Because the p-critical value of .05 is more forgiving, it would take a larger number of studies to reverse the findings of statistical significance within a fail-safe N analysis.

In Table 1, the fail-safe N analysis yielded an answer of 18,675. This means that in order to bring the hypothetical meta-analytic review's level of statistical significance down to the .05 level exactly, 18,675 nil-null result articles would be needed. According to Rosenthal's reasoning, this number markedly exceeds the 530 that would be considered the number needed to achieve stable results. Results such as these would be ideal for any researcher. Considering the fact that it would take over 18,675 nil-null articles to reverse the statistical significance findings of the hypothetical meta-analysis, the researchers should put a tremendous amount of faith in their results. This confidence

is due to the unlikely chance that 18,675 studies that found nil-null results were not included in the study because they failed to be published or because they were simply overlooked by the researchers in their literature review.

Orwin's Fail-Safe N Analysis Using Effect Size Measures

Using the theory developed by Rosenthal (1979), Orwin (1983) developed a fail-safe N that used effect size to determine the stability of the results within a meta-analysis. The formula is as follows:

$$N_{fs} = N_o (d_o - d_c) / d_c - d_{fs}$$

Where N_{fs} = the number of nil-null studies needed to reverse the statistically significant findings
N_o = the number of studies used in the meta-analysis
d_o = the mean effect size obtained for the meta-analysis
d_c = the criterion effect size value of the fail-safe studies
d_{fs} = the mean effect size of the fail-safe studies

The effect size statistic gives the amount of difference between treatment and control group means. By using this statistic, the researcher is able to get a more accurate read on the treatment results without being constrained by the reasoning involved in traditional significance testing.

One drawback to using the effect size measure is that there is no agreed upon criterion value. This is not true for p-critical levels, for which researchers historically have used either the .01 or .05 level to determine statistical significance. For effect size, Orwin (1983) suggested using Cohen's (1969) specifications of .2 as "small", .5 as "medium", and .8 as "large." Although this measure of effect size is not uniformly used throughout research studies, it is a sound method of analyzing the results reported.

It is also important to understand that the d_{fs} in the formula constructed by Orwin (1983) is usually assigned a value of 0. This is due to the fact that the file drawer studies

are hypothesized to have an effect size of 0. Although the majority of fail-safe N analyses attempt to find the number of nil-null studies needed to reverse the statistically significant results, this is not necessarily mandated within the formula. As Orwin stated, "a researcher may have reason to believe that the file drawer studies have a nonzero mean effect size, or he or she may wish to test a range of values around zero" (p. 158). This allows some freedom within the fail-safe analysis to substitute effect sizes not equal to zero within the "what if" analysis. Such a procedure would allow the researcher to explore the effects of varying effect sizes on any particular meta-analysis study and determine the amount of confidence that should be placed on the results.

In Table 2, the fail-safe N statistic determined that 11 articles were needed to lower the "medium" constant effect size to .5. This means that only 11 nil-null articles would be needed to change the effect size of .85 in the hypothetical study to the effect size of .5. This is a very drastic change considering the small number of nil-null studies that would be required to change the effect size. According to Rosenthal's formula, this number falls well below the 85 that would be needed to achieve stability in the results within this particular hypothetical study. Due to these factors, these results should be interpreted with considerable caution. Because it is conceivably possible that 11 nil-null articles could have been overlooked either because of their failure to be published or as a matter of oversight in the literature review, the results in this hypothetical study should be carefully interpreted due to the results of the fail-safe N analysis.

Limitations of the Fail-safe N

One of the shortcomings of the procedures outlined by Orwin (1983) and Rosenthal (1979) is the lack of a statistical model within these designs. Orwin stated,

"Although the statistic's utility as a heuristic device does not require one, the specification of a model or class of models describing its sampling distribution would be desirable" (p. 158). The development of a sample distribution for this particular statistical method would be very helpful in solidifying its importance in the statistical world today. Although this particular aspect of the fail-safe analysis has yet to be developed, there are definite benefits to the continued use of the statistics to help explain the characteristics contained within a meta-analysis.

Conclusion

The fail-safe N analysis can be considered a type of "what if" analysis that has been proposed by researchers such as Kieffer and Thompson (2000). The fail-safe N analysis previously outlined provides the researcher with several different options and benefits when interpreting the information from a meta-analysis. One of the benefits of this method is that the researcher is able to provide valuable information about the results of the study, rather than simply stating whether statistical significance was achieved or not. By providing this information in a clear and detailed manner, the stability of the research results can be conveyed to the researcher's audience.

It is important that researchers begin to take advantage of these methods to further evaluate their results. The shortcomings of reporting only whether statistical significance was achieved has been outlined in articles by Thompson (1992) and Cohen (1994). The use of "what if" analyses, such as the fail-safe N, can help to provide other professionals with important information concerning results and areas that should be researched further.

Formula 1

Rosenthal's Fail-safe N analysis using Z-scores:

$$X = [(SUM\ Z)^2 / G] - k$$

Where X = the number of studies needed to reverse the statistically significant findings
k = the number of studies combined
(SUM Z) = the sum of the Z scores for the individual studies
G = the Z-score that falls at the p-critical value being evaluated

Formula 2

Orwin's Fail-safe N analysis using Effect Sizes

$$N_{fs} = N_o (d_o - d_c) / d_c - d_{fs}$$

Where N_{fs} = the number of nil-null studies needed to reverse the statistically significant findings
N_o = the number of studies used in the meta-analysis
d_o = the mean effect size obtained for the meta-analysis
d_c = the criterion effect size value of the fail-safe studies
d_{fs} = the mean effect size of the fail-safe studies

References

Anderson, D.R., Burnham, K.P., & Thompson, W. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. Journal of Wildlife Management, 64, 912-923.

Brown, J.R. (1992). Detecting potential hucksterism in meta-analysis using a follow-up fail-safe test. Psychology in the Schools, 29, 179-184.

Carson, K.P., Schriesheim, C.A., & Kinicki, A.J. (1990). The usefulness of the "fail-safe" statistic in meta-analysis. Educational and Psychological Measurement, 50, 233-243.

Cohen, J. (1994). The earth is round (p , .05). American Psychologist, 49, 997-1003.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Eysenck, H.J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Glass, G.V. (1976). Primary, secondary, and meta-analysis research. Educational Researcher, 5, 3-8.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W.L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Orwin, R.G. (1983). A fail-safe N for effect size in meta-analysis. Journal of Educational Statistics, 8, 157-159.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22 (1), 2-5.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25 (2), 26-30.

Thompson, B., & Kieffer, K.M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. Research in the Schools, 7, (2), 3-10.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Table 1

Rosenthal's Fail-safe N Analysis

In 1999, 104 experiments examining the effects of Cognitive-Behavioral counseling on depression were summarized and statistically examined using a meta-analysis. The sum of the Z scores in this particular article was 175.76. Determine the number of articles necessary to reach significance at the .05 level.

$$X = [(SUM\ Z)^2 / G] - k$$

$$X = [(175.76)^2 / 1.645] - 104$$

$$X = [3089.58 / 1.645] - 104$$

$$X = 18779 - 104$$

$$X = 18{,}675$$

Where X = the number of studies needed to reverse the statistically significant findings
k = the number of studies combined in the meta-analysis
(SUM Z) = the sum of the Z scores for the individual studies
G = the Z-score that falls at the p-critical value being evaluated

Table 2

Orwin's Fail-safe N Analysis

In 1997, 15 articles examining the effects of Systematic Desensitization on panic attacks were summarized and statistically examined using a meta-analysis. The mean effect size for the studies in this particular article was .85. Find the fail-safe N for this meta-analysis using a "medium" constant effect size of .5.

$$N_{fs} = N_o (d_o - d_c) / (d_c - d_{fs})$$

$$N_{fs} = 15 (.85 - .5) / (.5 - 0.0)$$

$$N_{fs} = 15 (.35) / .5$$

$$N_{fs} = 5.25 / .5$$

$$N_{fs} = 10.5$$

Where N_{fs} = the number of nil-null studies needed to reverse the statistically significant findings
N_o = the number of studies used in the meta-analysis
d_o = the mean effect size obtained for the meta-analysis
d_c = the criterion effect size value of the fail-safe studies
d_{fs} = the mean effect size of the fail-safe studies